DOCUMENT RESUME

ED 476 148 TM 034 897

AUTHOR Wu, Brad C.

TITLE Scoring Multiple True False Items: A Comparison of Summed Scores and Response Pattern Scores at Item and Test Levels.

PUB DATE 2003-00-00

NOTE 40p.

PUB TYPE Reports - Research (143)

EDRS PRICE EDRS Price MF01/PC02 Plus Postage.

DESCRIPTORS *College Entrance Examinations; Foreign Countries; *High School Students; High Schools; *Item Response Theory; *Objective Tests; *Scoring

IDENTIFIERS Additive Models; Taiwan

ABSTRACT

The additive and response patterns scoring methods within and between multiple true-false (MTF) items were examined using data for 5,000 students for each of 2 years from the mathematics portion of the national college entrance examination in Taiwan. For additive scoring at item level, response to each option was scored dichotomously and added up to make an item clustered score, while at test level response to each item was scored dichotomously or polytomously by applying four methods, and then adding to sum of item score. For response patterns scoring the item response theory (IRT) ability estimated were estimated through the expected a posteriori procedure. The within-item IRT ability estimates were compared to the sum of the item scores at test level. Correlations between item clustered scores and within-item ability estimates were significant for all 10 items examined; correlations between sum of item scores and between-item ability estimates were also significant for all four scoring methods in two sets of tests. The results suggest that even at the risk of losing information, the use of item clustered scores and sum of item scores as estimates of the latent trait is reasonable, although the appropriateness of the item clustered scores should be examined prior to the test level estimation. The IRT ability estimates can be more informative when variation of discrimination parameters within items is large. The influence of item parameters on the IRT ability estimates was also discussed. (Contains 2 figures, 7 tables, and 39 references.) (Author/SLD)

Running Head: SCORING MULTIPLE TRUE FALSE ITEMS: A COMPARISON OF

Scoring Multiple True False Items: A Comparison of Summed Scores and Response Pattern Scores at Item and Test Levels

Brad C. Wu

University of Washington

Abstract

The additive and response patterns scoring methods within and between, multiple true-false (MTF) items were examined. For additive scoring at item level, response to each option was scored dichotomously and added up to an item clustered score, while at test level response to each item was scored either dichotomously or polytomously applying four methods and added up to a sum of item score. For response patterns scoring the IRT ability estimates were estimated through Expected *a posteriori* procedure. The within-item IRT ability estimates were compared to the clustered scores at item level and the between-item IRT ability estimates were compared to the sum of item scores at test level. Correlations between item clustered scores and within-item ability estimates were significant for all 10 items examined; correlations between sum of item scores and between-item ability estimates were also significant for all four scoring methods in two sets of test. The results suggest even at the risk of losing information, the use of item clustered scores and sum of item scores as estimates of the latent trait is reasonable. But the appropriateness of the item clustered scores should be examined prior to the test level estimation. The IRT ability estimates can be more informative when variation of discrimination parameters within items is large. The influence of the item parameters on the IRT ability estimates was also discussed.

Scoring Multiple True False Items: A Comparison of Sum Scores and Response Pattern Scores at Item and Test Levels

Despite many superior qualities such as higher reliability than the other item formats (Hills and Woods, 1978; Albanese et al, 1979; Albanese and Sabers, 1988; Mendelson et al, 1980; Frisbie and Sweeney, 1982; Kreiter and Frisbie, 1989) and more responses collected in a given time multiple-true-false (MTF) items possess, the application of this alternate format of the Multiple-Choice (MC) item has been limited. The major reason of its rarity can partly be attributed to the ambiguous status that complicates the scoring procedure and interpretation of the scores. Similar to the MC item, a typical MTF item consists of a question stem and a few options. The difference is that in responding to MTF items, examinees are asked to judge each option following the question stem as true or false instead of selecting only one correct option as required in MC items. Under such format, the scoring unit can either be an option or an item. In other words, the collection of responses can be scored for the item as a whole or response to each option can be scored separately. The dual scoring options also make it possible to score MTF items dichotomously or polytomously. While each option in the MTF item takes a form of an individual True False (TF) item, the content congruency among the options following each stem also brings the MTF item close to the format of Context Dependent Item Set (CDIS), in which options in an item are constructed as a subset, or a testlet (Wainer & Kiely, 1987; Wainer & Lewis, 1990).

When all the options are treated independently, as individual TF items, and scored dichotomously, guessing factor and local item dependency could seriously increase the error in estimating the latent trait. Clustered scoring is thus preferred because the sum

scores of within item TF responses are efficiently incorporated into one clustered score to provide estimates of performance at item level. This change helps to reduce the error due to the probability of guessing and serves as a general solution to the problem of local dependency (Yen, 1993). However, the efficacy of one clustered score for each item ignores the fact that a given score can result from different response patterns. For example, for an MTF item with 5 options, a clustered score of 3 could be a result of C_3^5 response patterns. Each pattern may reflect a different degree of the latent trait due to the varied characteristics of the options. Examinees who have same numbers of correct option but have different response patterns may have different degrees of latent trait due to the distinguished difficulty and discrimination of the options. Similarly, examinees could earn the same total score on a test based on different response patterns among items on the test. For a 20-item test, examinees with a score of 10 could demonstrate C_{10}^{20} response patterns. The question of whether to use sum of raw scores or response pattern scores in MC items, testlets, or tests combing different formats have been intensively discussed in the past, but rarely in relation to MTF items.

Sum of raw scores

In scoring MTF items, raw scores have always been used to estimate ability. Various linear arrangements on the raw scores can be made to yield different MTF scores. In a study by Albanese and Sabers (1988), different scoring methods based on raw scores of MTF items were examined for item and reliability analyses. When each option of the MTF item was treated as an independent true-false unit, dichotomous scoring applied to correct response to an option was scored as 1 and incorrect response to an option was scored as 0. When each MTF item with several options was treated as a unit,

each item could be scored either dichotomously or polytomously applying different scoring methods. For example, they used a clustered score for an item of several options instead of dichotomous scores for each option. Also, when only some of the options in an item were responded correctly, partial credit was given. These methods were developed to compensate for local dependency and guessing. Other methods such as correction-for-guessing was applied to discredit correct responses below chance level. The partial credit approach assigned credit only to total correct responses larger than chance level (half of the total options), and the correction-for-guessing approach subtracted credit as penalty for the incorrect responses. The combinations of these scoring methods led to the development of a) the multiple-response scoring, b) the count-for-2-options-correct scoring, c) the count-for-3-options-correct scoring, d) the correction-for-guessing scoring, e) the credit-for-any-correct scoring, and f) the separated-option scoring (Albanese and Sabers, 1988; Gross, 1978; Harasym, Norris, and Lorscheider, 1980; Sanderson, 1973).

Research suggests that correction for guessing does nothing but reduce the raw score to make the test seem more difficult (Hsu, Moss, and Khampalikit, 1989; Tsai and Suen, 1993). The approach of giving partial credit to partially correct responses, on the other hand, yields a higher raw score and also higher test score reliability when compared to a dichotomous scoring method such as the multiple-response scoring (Albanese, Sabers, 1988; Hsu, Moss, and Khampalikit, 1989). The MFT scoring methods discussed thus far are summarized in table 1.

Insert Table 1 Here

Response pattern scores

Response pattern scoring has been applied mostly in MC items and CR items but not in MTF items. Unlike the scoring methods using raw scores, response pattern scoring has usually been conducted with the application of various weighting schemes. Comparisons of implicit and explicit weighting have shown that the aim of achieving more detailed and reliable estimates of latent trait can be met by both. However, maximizing reliability by weighting may lead to lower validity (Rudner, 2001), therefore, weighting should be a rational process evaluating contributions and the trade-offs (Hennedy and Walstad, 1997).

Among the weighting schemes is the IRT procedure, which simultaneously calibrates all test items and estimates examinees' ability based on the item parameters. The consideration of the item parameters in ability estimation implicitly weighs each item (or option within an MTF item) and provides an IRT scaled score for each examinee. Under the IRT scheme, ability associated with each response pattern is usually estimated by the Maximum *a posteriori* (MAP) method or the Expected *a posteriori* (EAP) method. For the MAP, the mode of the joint likelihood derived from the product of corresponding trace lines and the $N(0,1)$ population distribution is calculated. For the EAP, the mean is used. The variation of the MAP or EAP estimates associated with a sum can be attributed to the variable parameters in the IRT models selected. IRT ability estimates derived from one-parameter logistic (Rasch) model are equivalent to summing the item scores because identical slope parameter and 0 guessing parameter are assumed. IRT ability estimates derived from the two-parameter logistic model are influenced by the item location (difficulty) and the slop (discrimination) parameters. For estimation based on the three-

parameter logistic model, item location and discrimination parameters affect the ability estimate (Lord, 1980, pp. 74-77). In other words, more discriminative items will have larger weights than less discriminative items.

While weighted scoring based on response patterns may lead to more detailed and reliable estimation of latent trait, raw scores have generally been used in scoring MC items, CR items, testlets, and tests of combined item formats. Besides the reason that raw scores are simple and convenient, Thissen (2001) argued that the difference in ability estimates resulting from IRT scaled scores and sum of raw scores is minor. The range of scaled scores around the sum scores is small because of a strong linear relationship between IRT scaled scores and the sum of raw scores for tests consisted of both MC and CR items.

For MTF items, the relationship of the IRT ability estimates and the sum of raw scores should be examined at both item and test levels. At the test level, the clustered scores of all MTF items add up to a total test score. Each total score is associated with various item response patterns and their corresponding IRT ability estimates. The variation of these ability estimates is determined by the parameters of each MTF item. Comparison of the IRT generated scaled scores and the total raw scores at the test level, however, relies on the appropriateness of the use of clustered scores in representing item performance. Unlike MC items, MTF items require judgments on several options and thus the patterns of judgment within MFT items should likewise be examined and compared to the corresponding clustered score.

The research questions of the study are:

1. Does the same type of linear relationship exist between sum of raw scores and IRT ability estimates based on response patterns in MTF items as in MC and CR items?
2. Is it appropriate to use clustered scores for an MTF item rather than separate scores for each option within an MTF item? Also, is it appropriate to use sum of item raw scores rather than weighted item response pattern scores to estimate latent trait?

The purpose of this study is to examine the relationship between the raw scores and the response pattern scores in MTF items, and the use of item clustered scores and sum of item scores as estimates of latent traits. This is done by first examining relationship between IRT ability estimates derived from response pattern scores within MTF items and item clustered scores, then between the ability estimates derived from different scoring methods and the sum of item scores stepwise.

Method

Instrument and Data

The data examined in this study are the MTF items from the Group I Mathematics test of the National College Entrance Examination (NCEE) held in Taiwan on July 2, 2000 and July 2, 2001. The Group I mathematics is the test for examinees who aim at majoring in Social Science, Art, Business, and other Humanistic Science, and thus the test places emphasis on mathematical knowledge and skills quite different from the Group II Mathematics test which is designed for those planning to major in the Natural Sciences. A total of 85,614 and 86,314 high school graduates took the Group I Mathematics test in 2000 and 2001 respectively. For each year, 5,000 examinees were

randomly selected from the population. Sample examinees were eliminated from analysis and ability estimation if any of the MTF item was left unanswered. The data screening resulted in sample examinees of 3,960 and 3,831 for year 2000 and 2001 respectively.

Each MTF item is followed by 5 options. This is similar to 25 TF items clustered into 5 testlets. The five MTF items for each year (see Appendixes A and B) cover the content in the standard Group I Mathematics curriculum from 10^{th} to 12^{th} grade, which includes Algebra, Geometry, Probability, and Statistics. A difference from the traditional MTF item is that the directions for the test indicated that at least one option following the item stem is true. This is sometimes called the Multiple Answer (MA) format where the number of possible response patterns becomes 2^k -1 (k is number of options). In a 5-option MTF item with at least one correct option, the number of response patterns is then 31 because "all false" pattern is excluded. The examinees' response patterns to all 5 items were collected for the analysis.

Insert Appendix A and Appendix B Here

Item Scoring

Each MTF item was scored twice and then compared. The first method was to score each option dichotomously as 0 (for incorrect response) or 1 (for correct response). The five option scores were added to create an item cluster score that ranges from 0 to 5. The item cluster score was mere summation of raw option scores. The second method was the IRT ability estimate, in which item score was estimated based on the response pattern to the five options. Prior to choosing the IRT method, the parameter estimations and item fit statistics were performed for each item. Model fitness was estimated using

the 2-parameter and the 3-parameter logistic models. The descriptive statistics for the residuals are shown in table 2. The standard residuals provide evidence that the item data were better fitted to the 3-parameter model. This is expected since guessing in TF item is inevitable. Thus EAP for each of the 31 response patterns to an item was estimated based on the 3-parameter logistic model.

Insert Table 2 Here

Test scoring

To score the whole MTF section that includes five MTF items, the score summation method and the ITR ability estimation from response patterns were both used again. In score summation, item cluster scores of each of the five items were added together to form a test score. Four different methods were applied here to produce item cluster scores.

1. The MR method gave 1 point to an item only when all the responses to the options were correct, 0 point was given otherwise. Therefore each item score was either 0 or 1 and the sum of item score (summation of five item scores) could be from 0 to 5.
2. The credit-for-any-correct scoring method, similar to the "MTF" scoring (Albanese and Sabers, 1988) gave 1 point to each correct response to an option. Thus item clustered score ranged from 0 to 5 and the sum of item score ranged from 0 to 25.
3. Similar to the count-for-3-option scoring, this method gave 1 point to an item when 3 of 5 responses were correct, 2 points when 4 of 5 responses were correct,

and 3 points when all 5 responses were correct. Credit was given only when the number of correct responses exceeds chance level (2.5 in this case). Since an item score had four categories (0 to 3) for this method, I named it the 4-category scoring method.

4. Similar to the count-for-4-option scoring, this method gave 1 point when 4 of 5 responses were correct, and 2 points when all the responses were correct. In this case, an item score had 3 categories (0 to 2) hence I named it the 3-category scoring method.

In sum, the four scoring methods used to score the MTF items are different in the number of scoring categories. Each item was scored using method with 2 score categories (MR scoring), 3 score categories (3-category-scoring), 4 score categories (4-category-scoring), and 6 score categories (credit-for-any-correct scoring). For each method, the 5 item scores were summed together.

To compare to the sum of item scores described above, the IRT ability estimates based on item score patterns were derived for each scoring method. For the MR scoring method, the IRT ability estimate was yielded from the pattern of five dichotomous scores (e.g. 01001). For the 3-category-scoring method, the IRT ability estimate was yielded from the pattern of five 3-category scores (e.g. 12201), and so forth for the other two methods. Each sum-of-item score had a corresponding IRT ability estimate.

The 2-parameter logistic model served as the basic model of parameter and IRT ability estimations for the MR method. The parameter and IRT ability estimations for the polytomous scoring methods (3-category-scoring method, 4-category-scoring method, and credit-for-any-correct scoring method) were performed based on Samejima's graded

response model. The graded response model was chosen because the amount of knowledge was assumed to have a positive relation with the score categories. The higher the item score, the better knowledge the examinee has. The location of each score category was expected to be in order. Also, the graded response model allows for different levels of discrimination for different items.

Correlation between scores

For each item, correlations were examined between the item cluster scores and the IRT ability estimates at the item level. For the two tests, correlations between sum of item scores and the IRT ability estimates at the test level under each scoring method were examined.

Test score reliability

For both tests, each of the 5 items was scored using five methods—the MR method, the 3-category method, the 4-category method, the credit-for-any-correct method, and the IRT ability estimate based on the response patterns. Test score reliability for the 5 methods were also examined.

Result

Item parameter estimation

Before the IRT ability estimation procedure, item parameters were generated from examinees' responses. Table 3 shows the item parameter estimates based on the 3-parameter logistic model. Parameters for each "option" were estimated because here each option was treated as a basic unit. The means of the discrimination parameters are .97 (S.D. = .50) and .70 (S.D. = .20) for the 2000 and 2001 test items respectively. The means of difficulty parameters were -.39 and -.65 with standard deviations of 1.15 and

.94 for the 2000 and 2001test items respectively. The guessing parameters averaged .48 and .38 with standard deviations of .08 and .03 respectively for the 2000 and 2001 MTF items.

Insert Table 3 Here

Parameters for each "item" based on the four scoring methods (MR, 3-category, 4-category, and credit-for-any-correct methods) were estimated and shown in table 6.

Insert Table 6 Here

Item scores comparison

The item cluster scores and the IRT ability (theta) estimates of all score patterns within an item are presented in table 4. For example, an examinee who responded correctly to only the second option on the first item of the 2000 mathematics test had a score pattern of 01000 that equals the item cluster score of 1, but an IRT item ability estimate of -.97. This score pattern (01000) for item 5 in the 2000 test would represent an examinee who left the item blank since options A, C, D, and E are all correct answers to the item stem. In other words, the examinee who left this item blank would have a score pattern of 0 for option A because option A is correct and should be chosen. He/she would have a score of 1 for option B because this option is incorrect and leaving it blank would earn him/her 1 score point. Similarly, zeros are given for the blank responses to option C, D and E. However, since the "01000" score response to item 5 in the 2000 test corresponds no response at all, it would not be scored using the IRT 3-parameter model. In fact, examinees with blank response to any item was eliminated from the analysis

because a blank is confounded by various situations such as believing all the options are incorrect (negligence of the direction that at least one option is correct), or not having enough time to respond.

Correlation coefficients of the item cluster scores and IRT ability estimates are shown in table 5. For the items, the correlations are all statistically significant ($p < .01$) and range from .78 to .98.

Insert Table 4 and Table 5 Here

Test scores comparison

When moving one hierarchical step up to the test level, each item was treated as a basic unit. The sum of item score and the IRT ability estimate for each test were derived and compared. The correlation coefficients between the sum of item scores and the IRT ability estimates are shown in table 7. Correlations of the IRT ability estimates with sum of item scores for each method range from .97 to .99 ($p < .01$).

Insert Table 7 Here

Test score reliability

The α coefficient of the 2000 test items is .61 when the credit-for-any-correct scoring method was used, .63 when either the 4-category or the 3-category scoring methods was used, .59 when the MR scoring method was used, and .62 when the IRT ability estimate was used. For the 2001 test items, the coefficient is .64 when the credit-for-any-correct scoring method was used, .65 when the 4-category method was used, .63

when the 3-category method was used, .58 when the MR method was used, and .63 when the IRT ability estimate was used. The score reliabilities do not vary much across the different scoring methods.

Discussion

As in the study of the MC and CR item scoring conducted by Thissen (2001), high correlations exist between the item cluster scores and the IRT ability estimates for each MTF items in the NCEE Mathematics tests. Such high correlations also exist between the sum of item scores and the IRT ability estimates for both 2000 and 2001 NCEE Mathematics tests. This is especially true at the test level. Although all the correlations between the item cluster scores and the IRT ability estimates are significant, the range of the IRT ability estimates corresponding to an item cluster score can be wide apart. For example, the minimum and maximum ability estimates for an item cluster score of 3 on item 2 in the 2000 mathematics test was −1.01 and 0 respectively. Plotting the item cluster scores against the IRT ability estimates (Figure 1 a) for item 1 in the 2000 mathematics test, linear relation between the two is not particularly overt even though the correlation is significant. The wider variation of the IRT ability estimates for middle range item cluster scores (2 to 4 points) can be clearly observed. This suggests that even with a strong linear relationship between the item cluster scores and the IRT ability estimates, the use of the item cluster scores will lose certain amount of information. Item 1 of the 2001 mathematics test (Figure 1b) represents a less varied and more linear relationship between the item cluster scores and the IRT ability estimates, though the IRT estimates corresponding to a cluster score could still be quite varied. Significant correlations provide evidence for strong relation between the two examinee

scores in a macroscopic manner, but microscopic analysis of ability estimation based on response pattern may show more variability in the IRT ability estimates corresponding a cluster score than we would expect and provide some lost information that we would want to capture.

Insert Figure 2 Here

One major reason for the variation of the IRT estimates corresponding to a cluster score is the varied discrimination power among the options in an MTF item. For item 1 of the 2000 mathematics test, the first two options are more discriminating than the other three options. Therefore, incorrect responses on these two options leads to a much lower IRT ability estimate (-1.55) than incorrect responses to any two of the other three options (-.55, -.55 and -.53). For examinees who respond to 3 options correctly on the item, the lowest IRT ability estimate is given to those who missed the first two options. The highest IRT ability estimate is assigned to examinees responding correctly to the three most discriminating options (1,2, and 5). The lowest IRT ability estimate (–1.55) is even lower than IRT ability estimates corresponding to only one correct response on the most discriminating options (-.97 and –1.11). The highest IRT ability estimate for examinees responding to 3 options correctly on the item (-.53) is greater than the IRT ability estimate corresponding to four correct responses when one of the two most discriminating options is missed (-.87 and -.83). The use of the item cluster score to represent all the response patterns in such case may lead to loss in information and erroneous estimation of ability at the item level.

The variation of the IRT ability estimates seems less problematic at the test level. A strong linear relationship between the sum of item scores and the IRT ability estimates can be seen regardless of the item scoring method used. Figure 3 shows clear linear relationship between the two types of scores for all item scoring methods. An average of 98% of sum of item score variance co-varies with the IRT ability estimates. Only limited information is lost using sum of item scores as estimates of ability. Thus sum of item scores under any scoring model seems more practical at test level than IRT ability estimates, assuming item cluster scores are appropriate as item score estimation.

Insert Figure 2 Here

Due to the variable discriminations and inevitable guessing in responding to options that complicate the ability estimation, IRT ability estimates are more likely to be inconsistent across the option response patterns at item level. Before test level estimation, it is worthwhile to examine the characteristics of the options within each MTF item to see how well the item cluster scores are going to represent the response patterns. This is important because both the additive and response-pattern scoring approaches at the test level are based on item cluster scores. Options with high guessing levels (>.5) should be identified and considered for revision. Also, dramatic differences of option discrimination within items can be a potential problem. These factors are the causes of variation among IRT ability estimation for different option response patterns.

To conclude, item cluster scores and sum of item scores are often used as ability estimates because they are simple to calculate and easy to interpret. In addition, they are usually closely related to the response pattern scores such as IRT ability estimates.

However, an IRT procedure's implicit weighting makes use of the item parameters to estimate ability based upon response patterns. The IRT ability estimates are thus thought as somewhat more efficient than the sum scores. Though the "extra information" might be fairly small, the IRT ability estimates provide details about how examinees perform and a convenient way to link alternate forms or constructs (Thissen, 2001).

References

Aiken, L. R. (1991). Detecting, understanding, and controlling for cheating on tests. Research in Higher, Education, 32, 725-736.

Albanese, M. A., Kent, T. H., & Whitney, D. R. (1979). A comparison of the difficulty, reliability, and multiple true-false items. Annual Conference on Research in Medical *Education, 16,* 105-110.

Albanese, M. A. (1982). Multiple-choice items with combination of correct responses. Evaluation and the Health Professions, *5*, 212-228.

Albanese, M. A., & Sabers, D. L. (1988). Multiple true-false items: A study of interitem correlations, scoring alternatives, and reliability estimation. Journal of Educational Measurement, 25, 111-123.

Bliss, L. B., & Mueller, R. J. (1987). Assessing study behaviors of collage students. Journal of Developmental Education, *11*, 14-18.

Britten, B. K., Glynn, S. M., Meyer, B. J. F., & Penland, M. S. (1982). Effects of test structure on use of cognitive capacity during reading. Journal of Educational Psychology, 74, 51-61.

Coombs, P. L., Milholland, J. E., & Womer, F. B. (1956). The assessment of partial Knowledge. Educational and Psychological Measurement, 16, 13-17.

Downing, S. M. (1995). Item type and cognitive ability measured: The validity evidence for multiple true-false items in medical specialty certification. Applied Measurement in Education, 8, 187-197.

Dressel, P. L., & Schmid, J. (1953). Some modifications of the multiple-choice item. Educational and Psychological Measurement, 13, 574-595.

Duncan, G. T., & Milton, E. O. (1986). Multiple-answer multiple-choice test items: Responding and scoring through Bayes and Minimax strategies. Psychometrika, 43, 43-57.

Frary, R. B. (1982). A simulation study of reliability and validity of multiple-choice test scores under six responses scoring modes. Journal of Educational Statistics, 7, 333-51.

Frisbie, D. A. (1974). The effect of item format on reliability and validity: a study of multiple choice and true-false achievement tests. Educational and Psychological Measurement, 34, 885-892.

Frisbie, D. A., & Sweeney, D. C. (1982). The relative merits of multiple true-false achievement tests. Journal of Educational Measurement, *19*, 29-35.

Frisbie, D. A. (1992). The multiple true-false item format: A status review. Educational Measurement: Issues and Practice, 11, 21-26.

Gross, L. J. (1979). Considerations in scoring multiple true-false tests. Health Professions Education Bulletin, 7, 26-30.

Haladyna, T. (1992). The effectiveness of several multiple-choice formats. Applied Measurement in Education, 5(1), 73-78.

Harasym, P. H. (1993). Negation in stems of single response multiple-choice items: An overestimation of student ability. Evaluation and the Health Professions, 16, 342-57.

Harasym, P. H. (1992). Evaluation of negation in stems of multiple-choice items. Evaluation and the Health Professions, 15, 198-220.

Hills, G. C., & Woods, G. T. (1974). Multiple true false questions. Education in Chemistry, *11*, 86-87.

Hsu, T., & Khampalikit, C. (1982). Application of item response theory to non-tryout Tests constructed for college admission testing. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Hsu, T., Moss, P. A., & Khampalikit, C. (1993). Merits of multiple answer items as evaluated by using six scoring formulas. Journal of Experimental Education, 69, 52-158.

Huberty, C. J., Julian, M. W. (1995). An ad hoc analysis strategy with missing data. Journal of Experimental Education, 83, 333-42.

Huntley, R. M., Plake, B. S. (1988, April). An investigation of multiple-response option multiple-choice items: Items performance and processing demands. Paper presented at the Annual Meeting of the American Research Association, New Orleans, LA.

Israel, G. D., & Taylor, C. L. (1990). Can response order bias evaluations? Evaluation and Program Planning, 13, 365-71.

Kenny, D. A., Zautra, A. (1995). The trait state error model for multiwave data. Journal of Consulting and Clinical Psychology. 83, 52-58.

Kolstad, R. K., Wagner, M. J., Kolstat, R. A., & Miller, E. G. (1983). The failure of distractors on complex multiple-choice items to prevent guessing. Educational Research Quarterly, 8, 44-50.

Kreiter, C. D., & Frisbie, D. A. (1989). Effectiveness of multiple true-false items. Applied Measurement in Education and Psychology, 2, 207-216.

Lee, G., Brennan, R., & Frisbie, D. (2000). Incorporating the testlet concept in test score analyses. Educational Measurement: Issues and Practice, 19, 9-14.

Lin, S. S. (1993). Fitting item response theory models to the College Entrance Examination of Taiwan. Unpublished doctoral dissertation, University of Oregon, Eugene.

Mendelson, M. A., Hardin, J. H., and Canady, S. D. (1980). The effect of format on the difficulty of multiple-completion test items. Paper presented in at the Annual Meeting of the National Council on Measurement in Education, Boston.

Namy, E. A. (1995) Comparison of paired multiple response items and multiple choice items. Psychological Reports, 25, 583-85.

Rudner, L. (2001). Informed test component weighting. Educational Measurement: Issues and Practice, 20, 16-19.

Schriesheim, C., & Schriesheim, J. (1994). Development and empirical verification of new response categories to increase the validity of multiple response alternative questionnaires. Educational and Psychological Measurement, 34, 877-84.

Stape, C. J. (1995). Techniques for developing higher-level objectives test questions. Performance and Instruction, 34, 31-34.

Thissen, D., & Wainer H. (2001). Test scoring. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Tsai, F., & Suen, H. (1993). A brief report on a comparison of six scoring methods for multiple true and-false items. Educational and Psychological Measurement, 53, 399-404.

Wang, C., & Terry, A. (1994, April). An examination of response dependency when there is more than one correct answer. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Los Angeles.

Wesman, A. G. (1971). Writing the test item. Washington, DC: American Council on Education.

Woodruff, D. J., & Feldt, L. S. (1985). Comparing the alpha coefficients of several tests when all tests are administered to the same sample. (ACT Technical Bulletin, No. 48). Iowa City, IA: ACT.

Table 1

Summary of Scoring Methods for a 5-option MTF Item

Basis of Unit	Correction for Guessing	Above Chance Level Credit[1]	Credit Given to Any Correct Response	Scoring Formula	Scoring Method
Item-based (Test level)	no	no	no	$F(i)=1$ if i=k $F(i)=0$ otherwise	Multiple Response Scoring
	no	yes	no	$F(i)=0$ if i=0,1,or2 $F(i)=\frac{i-2}{k-2}$ otherwise	Count-for-3-options-correct Scoring
	no	yes	no	$F(i)=0$ if i=0,1,2,or 3 $F(i)=\frac{i-3}{k-3}$ otherwise	Count-for-4-options-correct Scoring
	yes	no	yes	$F(i)=\frac{i-(k-i)}{k}$ where $k \geq i \geq 0$	Correction for Guessing Scoring
	no	no	yes	$F(i)=\frac{i}{k}$ where $k \geq i \geq 0$	MTF scoring
Option-based (Item level)	no	no	no	$F(i)=1$ if correct $F(i)=0$ if incorrect	Separated option Scoring

[1] In a 5-option MTF item, the chance level is 2.5. An examinee thus needs at least 3 correct responses to receive credit.
k = number of options in the MTF item.
i = number of correct responses.

Table 2

Estimates of Standard Residuals at Item Level Fitting the 2-parameter and the 3-parameter Logistic Models

Test and Model (sample size)	Standard Residuals			
	Range		Mean	S. D.
	Minimum	Maximum		
Math 2000 2-parameter (3960)	.64	3.38	1.5	.77
Math 2000 3-parameter (3960)	.42	1.59	1.08	.28
Math2001 2-parameter (3831)	.39	2.15	1.01	.40
Math2001 3-parameter (3831)	.34	.44	.38	.03

Table 3
Parameter Estimates of the 10 MTF Items Treating Each Option as a TF Item

Parameter		Year 2000			Year 2001		
		a	b	c	a	b	c
Item 1	Option 1	1.29	-0.99	.39	.88	1.10	.44
	Option 2	1.24	-1.01	.41	.68	-0.30	.36
	Option 3	.84	-0.52	.50	.60	-0.30	.36
	Option 4	.83	-0.33	.42	.71	-0.81	.34
	Option 5	.86	0.37	.59	.55	0.66	.39
Item 2	Option 1	.67	-1.03	.45	1.11	-0.65	.37
	Option 2	.71	1.35	.49	.73	-0.67	.37
	Option 3	.64	-0.77	.43	.44	-0.56	.39
	Option 4	.79	-0.40	.38	1.23	0.60	.39
	Option 5	.39	-0.27	.52	.76	-0.48	.36
Item 3	Option 1	.93	1.77	.51	.76	-1.65	.37
	Option 2	.91	0.29	.59	.53	-1.22	.37
	Option 3	.90	0.74	.61	.83	0.04	.43
	Option 4	1.28	1.29	.68	.79	-1.75	.36
	Option 5	.40	-1.11	.53	1.04	0.46	.39
Item 4	Option 1	1.63	0.67	.53	.67	-1.41	.39
	Option 2	1.49	-0.31	.42	.66	-1.06	.34
	Option 3	2.24	-0.23	.40	.56	-1.15	.35
	Option 4	2.16	-0.28	.42	.62	-2.50	.37
	Option 5	1.15	0.58	.44	.72	-1.44	.39
Item 5	Option 1	1.01	-0.56	.35	.76	-1.20	.35
	Option 2	.48	-2.49	.50	.49	-1.51	.37
	Option 3	.43	-1.30	.50	.42	-0.50	.37
	Option 4	.68	-2.29	.46	.60	1.31	.44
	Option 5	.40	-2.81	.50	.46	-1.19	.37

a=discrimination b=location (difficulty) c=guessing (lower asymptote)

Table 4

Item Clustered Scores and IRT Theta Estimates for All Possible Option Score Patterns Within an MTF Item in the 2000 and 2001 NCEE Group I Mathematics Test

Possible Option Score Patterns for Each Item	Item Clustered Scores	IRT Theta Estimates for Examinees on Each Item									
		Year 2000 Test					Year 2001 Test				
		Item 1	Item 2	Item 3	Item 4	Item 5	Item 1	Item 2	Item 3	Item 4	Item 5
00000	.00	-1.60	*	-1.51	-1.60	-2.12	-1.38	-1.50	-1.44	-1.75	-1.29
10000	1.00	-1.11	-1.24	-1.51	-1.60	-1.89	*	-.56	-1.43	-1.56	-1.29
01000	1.00	-.97	-1.25	-1.24	-1.17	*	-.79	-1.30	-1.21	-1.75	-1.22
00100	1.00	-1.60	-.81	-1.02	-1.29	-2.08	-1.03	-1.50	-1.14	-1.75	-.84
00010	1.00	-1.60	-1.07	-1.51	-.51	-1.57	-1.20	*	-.97	-1.59	*
00001	1.00	*	-1.04	-.98	-1.46	-2.11	-1.37	-1.50	-1.44	-1.47	-.70
11000	2.00	.55	-1.24	-1.24	-1.17	-1.20	-.65	-.51	-1.20	-1.55	-1.22
10100	2.00	-1.11	-.69	-1.02	-1.29	-1.82	-.83	-.55	-.81	-1.55	-.84
10010	2.00	-1.10	-.95	-1.51	-.51	-1.38	-.90	-.52	-.96	-1.14	-1.44
10001	2.00	-.98	-1.02	-.98	*	-1.88	-1.13	-.37	-1.43	*	-.63
01100	2.00	-.97	-.80	-.65	-.72	-1.22	-.56	-1.30	-.91	-1.75	-.77
01010	2.00	-.96	-1.07	-1.24	-.50	-.91	-.38	-1.02	-.83	-1.59	-1.36
01001	2.00	-.86	-1.04	-.56	-.96	-1.36	-.76	-1.30	-1.21	-1.46	-.59
00110	2.00	-1.60	-.42	-1.02	-.39	-1.43	-.70	-1.25	-.72	-1.59	-.96
00101	2.00	-1.55	-.64	-.34	-1.02	-2.06	-1.01	-1.49	*	-1.46	-.20
00011	2.00	-1.55	-.83	-.98	-.50	-1.54	-1.14	-1.24	-.97	-.96	-.80
11100	3.00	-.55	-.68	-.65	-.72	-.99	-.44	-.49	-.44	-1.55	-.75
11010	3.00	-.55	-.93	-1.24	-.50	-.77	-.14	-.48	-.82	-1.13	-1.36
11001	3.00	-.53	-1.01	-.56	-.96	-1.14	-.60	-.05	-1.20	-.58	-.46
10110	3.00	-1.05	.00	-1.02	-.26	-1.20	-.40	-.50	-.09	-1.13	-.95
10101	3.00	-.98	-.46	-.34	-1.02	-1.79	-.80	-.13	-.23	-.58	.13
10011	3.00	-.98	-.64	-.98	-.50	-1.34	-.78	-.18	-.96	-.51	-.74
01110	3.00	-.95	-.41	*	-.05	-.72	.05	-1.02	-.62	-1.59	-.87
01101	3.00	-.86	-.64	.19	-.61	-1.03	-.49	-1.29	-.89	-1.45	.01
01011	3.00	-.86	-.83	-.55	-.50	-.81	-.12	-1.01	-.82	-.95	-.68
00111	3.00	-1.55	-.19	-.32	-.23	-1.00	-.51	-1.24	-.68	-.95	-.24
11110	4.00	.39	.34	-.64	.32	-1.34	.24	-.44	.10	-1.12	-.85
11101	4.00	-.53	-.44	.19	-.61	-.65	-.35	.42	.19	-.17	.62
11011	4.00	-.53	-.60	-.53	-.50	-.62	.22	.16	-.81	-.51	-.56
10111	4.00	-.87	.24	-.27	.04	-1.02	-.12	.17	.57	-.51	.09
01111	4.00	-.83	-.18	.34	.20	.00	.55	-1.00	-.56	-.92	-.03
11111	5.00	.71	.76	.83	.85	.43	.90	.82	.82	.60	.59

* Indicates blank response to the item, which was eliminated from estimation procedure. The theta for the blank response was not estimated under zero frequency.

Table 5

Correlations Coefficients of Sampled Examinees' Item Clustered Scores and IRT Ability Estimates on the 10 MTF Items

	Year 2000 Test					Year 2001 Test				
Item	1	2	3	4	5	1	2	3	4	5
γ	.92**	.92**	.78**	.94**	.98**	.96**	.97**	.98**	.94**	.85**

** $p < .01$

Table 6

Parameter Estimates of the Five MTF Items for the Four Scoring Methods Based on 3-parameter Logistic Model and Graded Response Model

Scoring Method		2000 Test										2001 Test									
		Item 1		Item 2		Item 3		Item 4		Item 5		Item 1		Item 2		Item 3		Item 4		Item 5	
		a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b
Multiple Response		.91	.00	.56	.62	.97	.29	.77	-.18	.65	.49	.83	.66	1.15	.37	.98	.19	.44	.49	.65	.58
3 Category	L*		-.43		.45		.13		-1.13		.17		.54		.09		-.16		-1.33		.08
	1	.90	.38	.39	1.02	.84	.42	.76	.77	.78	.66	.75	.70	.99	.56	.87	.50	.44	.47	.60	1.0
	2		-.38		-1.02		-.42		-.77		-.66		-.70		-.56		-.50		-.47		-1
4 Category	L*		-.85		-.64		-.35		-1.71		-.48		-.04		-.36		-.78		-2.02		-.63
	1	.89	.83	.37	2.22	.73	.94	.78	1.23	.83	1.29	.74	1.16	.91	.88	.89	1.23	.46	1.49	.64	1.42
	2		-.03		-.05		-.01		.14		-.01		.12		.16		-.13		-.3		.23
	3		-.8		-2.17		-.93		-.137		-1.28		-.128		-1.04		-.111		-.119		-.166
Credit for Any Correct	L*		-1.91		-1.15		-3.52		-1.42		-3		-1.19		-1.41		-1.86		0		-2.16
	1	.87	2	.35	2.07	.68	6.21	.8	2.41	.86	2.61	.73	2.59	.87	2.43	.88	2.55	.12	3.14	.63	3.21
	2		1.17		.57		2.34		.74		1.33		1.13		.95		.93		2.02		1.65
	3		-.22		-.7		-.52		-.05		-.1		-.06		-.21		.9		1.04		-.17
	4		-1.09		-1.94		-2.9		-1.06		-1.18		-1.12		-.97		-1.28		-2.05		-1.38
	5		-1.87				-5.13		-2.04		-2.66		-2.54		-2.2		-2.28		-4.14		-3.31

* L stands for location parameter, which is followed by category parameters.

Table 7

Correlation Coefficients of the Sampled Examinees' Sum Scores and IRT Ability Estimates for 4 Scoring Methods

Scoring Method	2000 Test	2001 Test
MR	.99**	.99**
3-category	.99**	.99**
4-category	.98**	.98**
Credit-for-any-correct	.97**	.97**

** $p < .01$

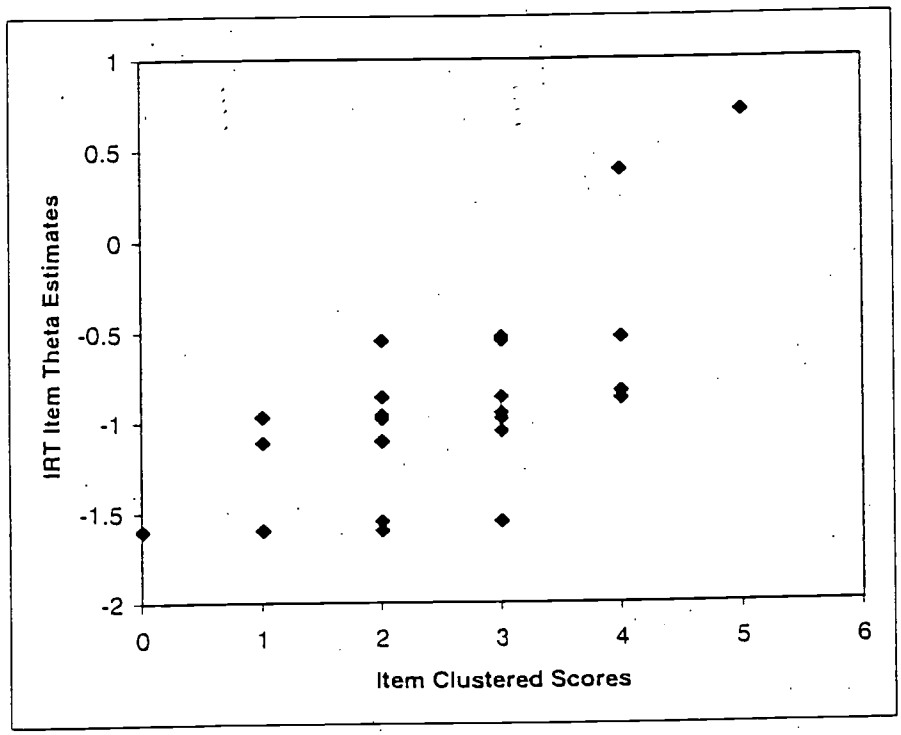
Figure caption

Figure 1. Scatter Plots of the Item Clustered Scores Against the IRT Ability Estimates for 2 MTF Items.

Figure 2. Scatter Plots of the Sum of Item Scores Against the IRT Ability Estimates for 4 Scoring Methods on MTF Items in 2000 Test.

Figure 1

a. MTF Item 1 (Year 2000)



b. MTF Item 1 (Year 2001)

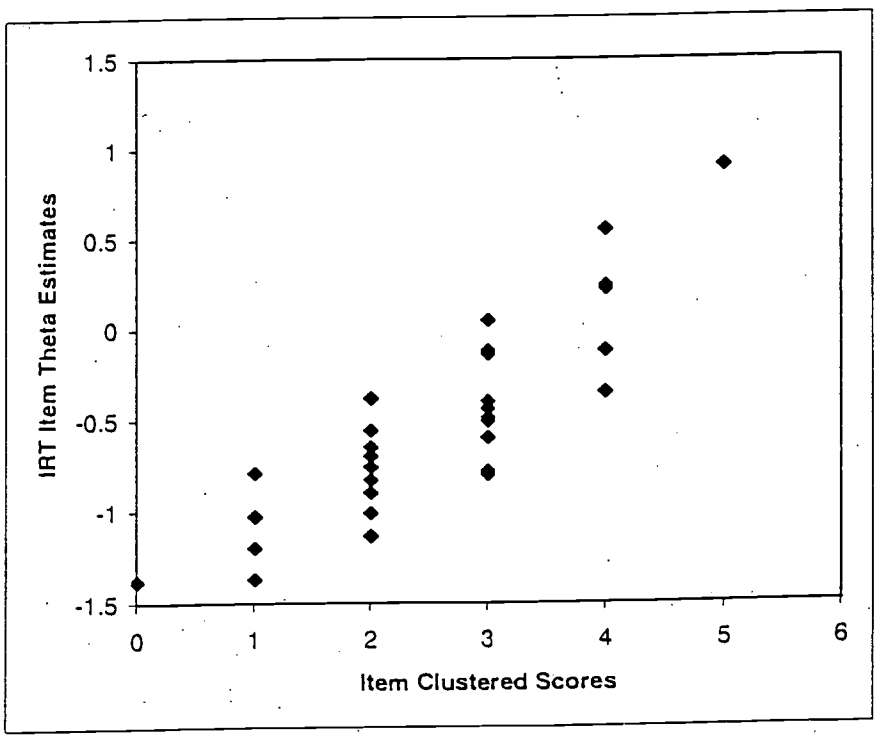
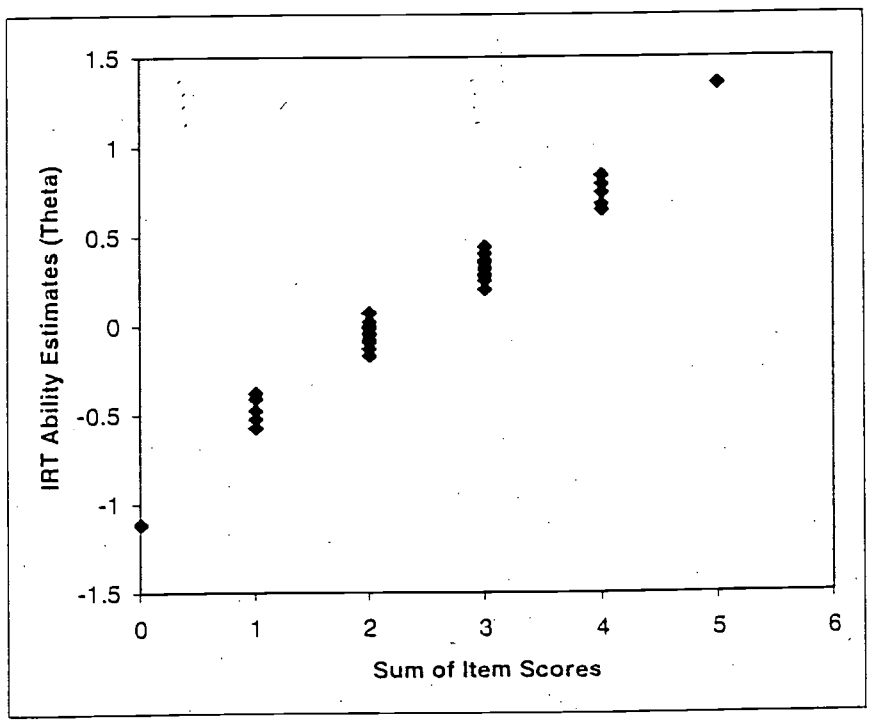
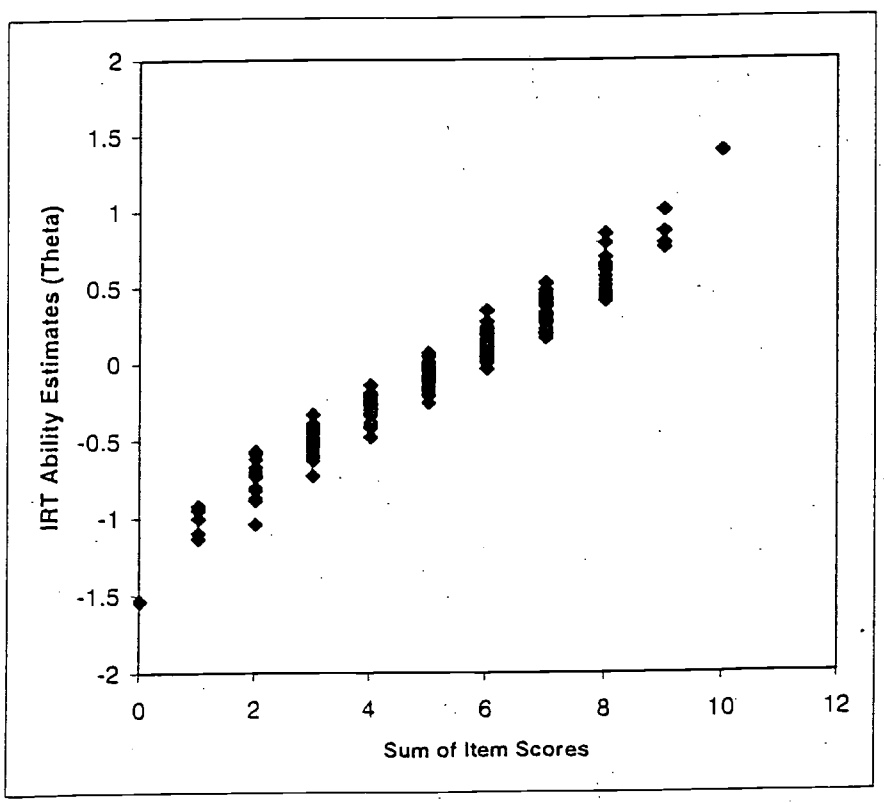


Figure 2

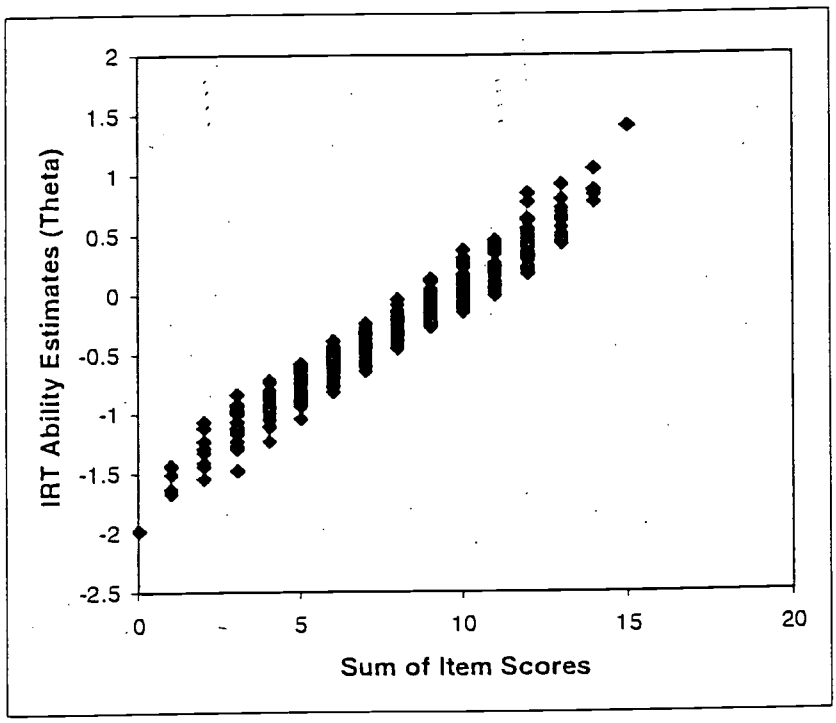
a. The MR Scoring



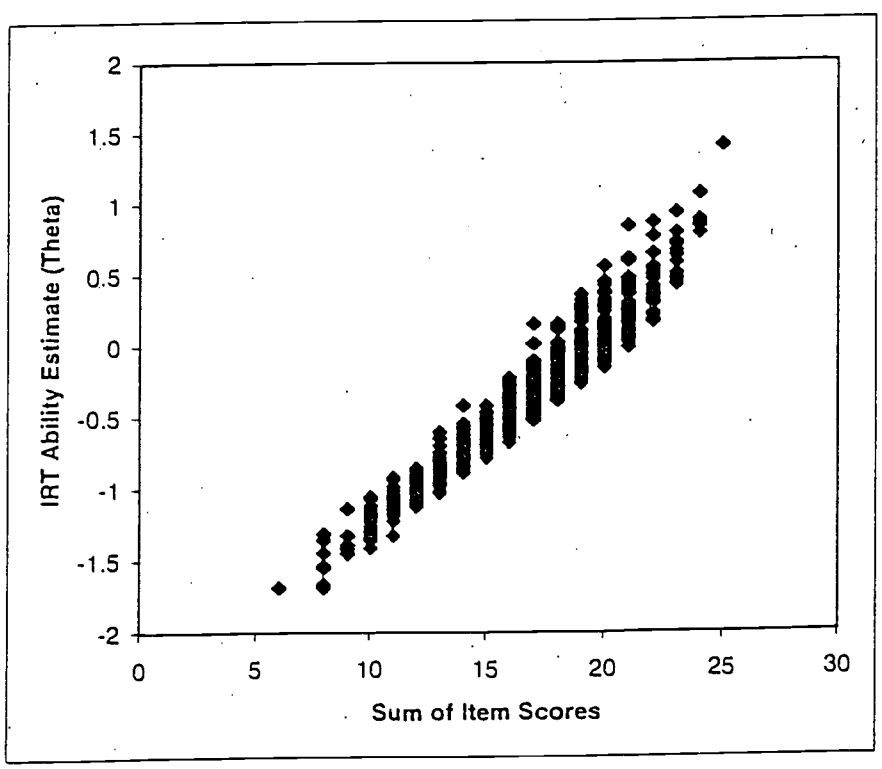
b. The 3-category scoring



c. The 4 category scoring



d. The credit-for-ant-correct Scoring



Appendix A

The MFT items of the 2000 NCEE Mathematics Test

Item 1

Which of the following statements are true about the equation $f(x)=x^4-15$?

a. There is one real solution between 1 and 2. *

b. There is one real solution between –2 and –1. *

c. There is no real solution larger than 2. *

d. There is no real solution smaller than –2. *

e. There are 4 real solutions.

Item 2

Which of the following are true?

a. $\frac{2^{10}+2^{20}}{2} > \sqrt{2^{10}\times 2^{20}}$ *

b. $\frac{\left(\frac{1}{2}\right)^{10}+\left(\frac{1}{2}\right)^{20}}{2} > \sqrt{\left(\frac{1}{2}\right)^{10}\times\left(\frac{1}{2}\right)^{20}}$ *

c. $\sqrt{10}+\sqrt{20} > \sqrt{30}$ *

d. $\log 10+\log 20 > \log 30$ *

e. $\frac{10^2+20^2}{2} > \left(\frac{10+20}{2}\right)^2$ *

Item 3

Which line(s) in the cubic ABCD-A'B'C'D' as seen below can be on the same plane with $\overline{A'B}$?

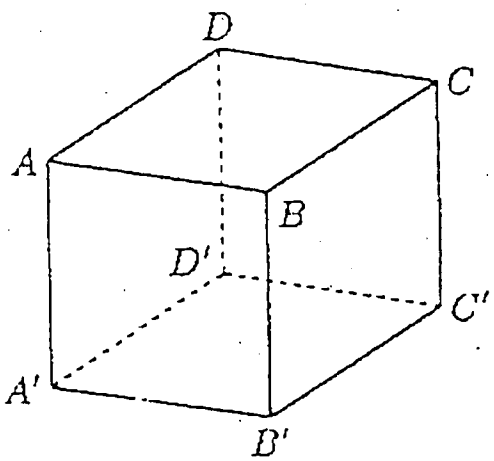
a. $\overline{BC'}$ *

b. $\overline{AC}$

c. $\overline{DB'}$

d. $\overline{DD'}$

e. $\overline{CD'}$ *



Item 4

The outer region of the parabola C: $y = -4x^2 + 9$ is divided into two areas by the y-axis. Which point(s) are located in the same side as the parabola's focal point?

a. (1.5,2)

b. (1,4) *

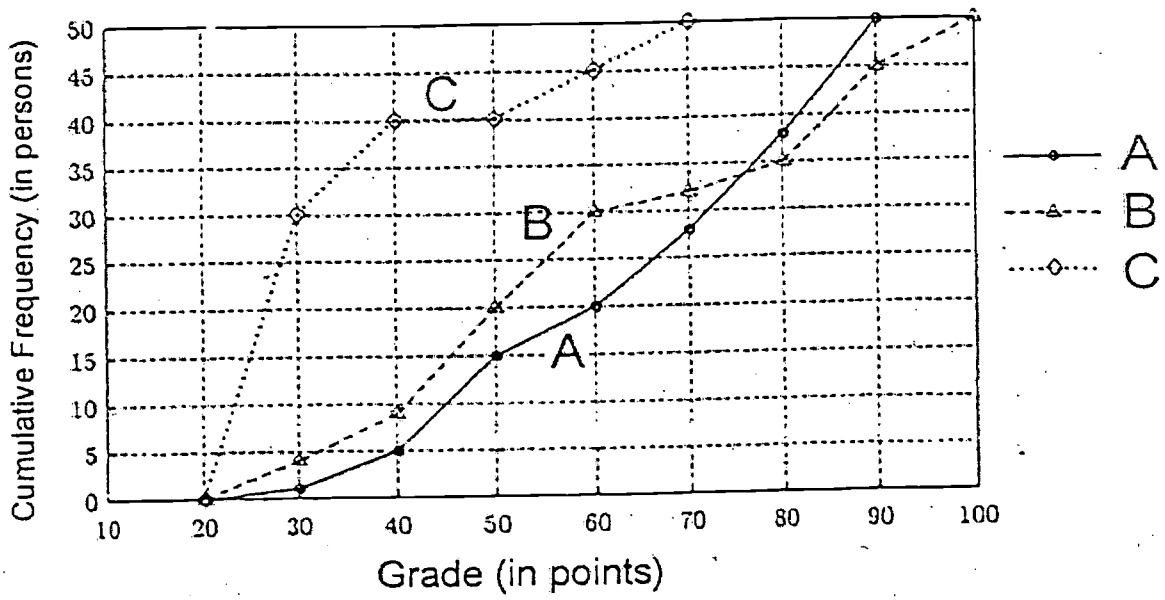
c. (–0.5,7) *

d. (0.5,7) *

e. (0,9)

Item 5

The plot below shows the cumulative frequencies of the Mathematics midterm for three 10^{th} grade classes. To pass the test, one should at least get 60 points.



Which of the following statement(s) is (are) true?

a. The Class "A" has the highest median. *

b. The Class "C" has the most passing examinees. *

c. The class "B" has the highest frequency of examinees that score 80 or higher.

d. The Class "C" has the lowest mean. *

e. The highest score is in the class "B." *

* indicates the correct option.

Appendix B

The MFT items of the 2001 NCEE Mathematics Test

Item 1

The real number $x = \frac{\sqrt{5}-1}{2}$, which of the following option has a value that is equivalent to x?

A. .62

B. $\frac{1}{x} - 1$ *

C. $1-x^2$ *

D. $\frac{1}{1+x}$ *

E. $1 - x + x^2 - x^3 + ... + (-1)^n x^n$, where $n \to \infty$ *

Item 2

If a, b, c are real numbers, and for $f(x)=ax^2 + bx + c$, $f(-1) = -3$, $f(3) = -1$, $b^2 - 4ac < 0$, then

A. $a < 0$ *

B. $c < 0$ *

C. $f(0) < f(1)$ *

D. $f(4) < f(5)$

E. $f(-3) < f(-2)$ *

Item 3

With which of the following condition(s) known can we find the equation of an ellipse on a 2-dimentional plane?

A. The locations of the four pinnacles of the ellipse. *

B. The locations of the two focal points and one point on the ellipse. *

C. The length of the short and long axis of the ellipse.

D. The locations of the two focal points and the length of the long axis of the ellipse. *

E. The location of the center and the ratio of the length of the long axis to the short axis of the ellipse.

Item 4

The mathematician and philosopher Galileo infuriated the Roman church when the "Dialogue" was published in 1632. He was then sentenced to life time imprison at the age of 70 and died in jail at the age of 78. The time between the publication of the "dialogue" and his death was said to be the darkest 10 years in his life. Galileo invented the 10x telescope in his early age and found the satellite "Europa" of Jupiter the year following the invention of the 10x telescope. The golden age of Galileo spanned from the invention of the telescope to the publication of the "Dialogue." The length of the golden age (in years) is half the number of his age when he found Europa. Based on the historical facts above, determine which of the following statement(s) is (are) true.

A. Galileo was born in 1566.

B. Galileo invented the 10x telescope at the age of 45. *

C. Galileo found the satellite "Europa" in 1610. *

D. The "Dialogue" was published when Galileo was 68 of age. *

E. Galileo died in 1644.

Item 5

Statistics of number of hours spent on operating computer per week for a class of 40 students is Shown below:

Mean hours spent on computer	8.3 hours
Standard Deviation	2.1 hours
The first quartile (Q1)	7.0 hours
The third quartile (Q3)	10.0 hours

Based on the statistics provided, determine which of the following statement(s) is (are) true?

A. The quartile deviation is 1.5 hours. *

B. The median is between 7.0 hours and 10.0 hours. *

C. There are 10 students who operate computer more than 10.0 hours a week. *

D. The longest time spent on operating computer is 12.5 hours $(8.3 + 2 \times 2.1)$ per week.

E. There are 20 students who spend 7 to 10 hours per week operating computer. *

* indicates the correct option.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

TM034897

I. DOCUMENT IDENTIFICATION:

Title: Scoring Multiple True False Items: A Comparison of Summed Scores and Response Pattern Scores at Item and Test Levels	
Author(s): Brad Ching-Chao Wu	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

X

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature:	Printed Name/Position/Title: Brad Ching-Chao Wu
Organization/Address: University of Washington	Telephone: 206-363-6901 FAX: 206-363-6901
	E-Mail Address: [illegible]@u.washington.edu Date: 4/23/03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org